Learning from Imabalanced Data
An Application to Bank Fraud Detection

Guillaume Metzler

Univ. Lyon, UJM-Saint-Etienne, CNRS, Institut d'Optique Graduate School, Laboratoire Hubert Curien UMR 5516, F-42023, SAINT-ETIENNE, France
Blitz Business Services, VILLEFONTAINE, France

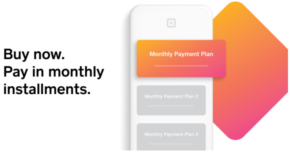
25 September 2019

Marianne CLAUSEL	Professeure, Université de Lorraine	Rapporteure
Marc TOMMASI	Professeur, Université de Lille	Rapporteur
Yves GRANDVALET	Professeur, Université de Technologie de Compiègne	Examinateur
Élisa FROMONT	Professeure, Université de Rennes I	Co-encadrante
Amaury HABRARD	Professeur, Université de Saint-Étienne	Co-encadrant
Marc SEBBAN	Professeur, Université de Saint-Étienne	Directeur
Xavier BADICHE	Président Directeur Général de Blitz Business Services	Blitz BS
Brahim BELKASMI	Ingénieur R&D, Blitz Business Services	Blitz BS

Context of Thesis

Blitz company

Blitz activities

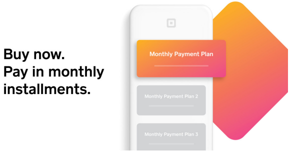


Payment facilities

Smooth checkout flow

Securing cheque transactions

Blitz activities



Payment facilities

Smooth checkout flow

Securing cheque transactions

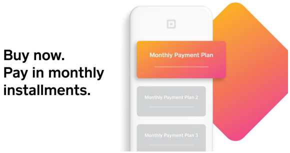
Other activities:

- Assistance with PV management
- Assistance in staff management

Context of Thesis

Blitz company

Blitz activities



Payment facilities

Smooth checkout flow

Securing cheque transactions

Other activities:

- Assistance with PV management
- Assistance in staff management

→ **This work Focus on the topic of securing cheque transactions ...**

What is cheque fraud ?

- Unpaid cheque (no money on bank account)
- False cheque
  - Not the real identity
  - Incorrect number series in the CMC7

Context of Thesis
Check fraud detection

What is cheque fraud ?

- Unpaid cheque (no money on bank account)
- False cheque

  Not the real identity

  Incorrect number series in the CMC7

Some statistics :

10 months of transactions (03/20/2016 to 10/21/2016)

- around 3.2 millions of transactions
- for 195 millions of euros
- 20 000 are frauds or unpaid (0.6%)
- represent 2 millions of euros (1.1%)

What is cheque fraud ?

- Unpaid cheque (no money on bank account)
- False cheque

  Not the real identity

  Incorrect number series in the CMC7

Some statistics :

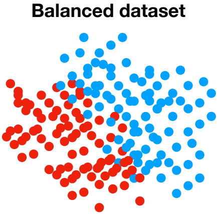
10 months of transactions (03/20/2016 to 10/21/2016)

- around 3.2 millions of transactions
- for 195 millions of euros
- 20 000 are frauds or unpaid (0.6%)
- represent 2 millions of euros (1.1%)

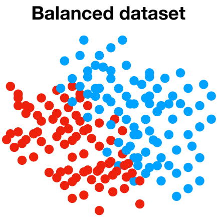
... more precisely on the topic of learning from imbalanced data

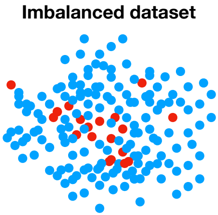
Outline

1. **Introduction on Learning From Imbalanced Data**
2. **A Geometrical Approach based on the Distance to Positives**
   2.1 Building Risky Areas
   ME^2 : *"Learning Maximum Excluding Ellipsoids from Imbalanced Data with Theoretical Guaranties"*
   2.2 An Adjusted Version Nearest Neighbor Algorithm
   $\gamma - k$*-NN : "An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data"*
3. **An Approach based on Cost-Sensitive Learning**
   3.1 Optimizing F-measure by Cost-Sensitive Classification
   CONE*: "From Cost-Sensitive Classification to Tight F-Measure Bounds"*
   3.2 Improving the Benefits of Mass Distribution
   "Tree-based Cost-Sensitive Methods for Fraud Detection in Imbalanced Data"
4. **Conclusion and Perspectives**



Positives $\simeq$ Negatives





Positives $\simeq$ Negatives

Positives $\ll$ Negatives

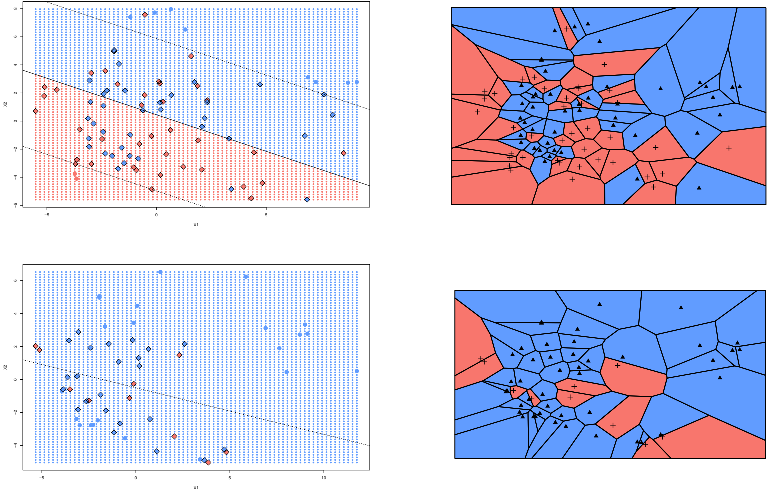
Minimizing a surrogate of $\frac{1}{m}\sum_{i=1}^{m} 1_{\{\hat{y}_i \neq y_i\}}$ leads to:

focus on both classes

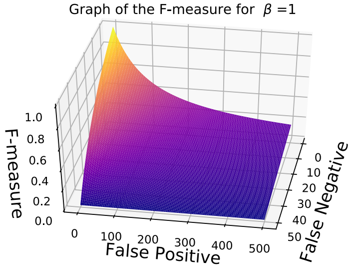
focus on majority examples

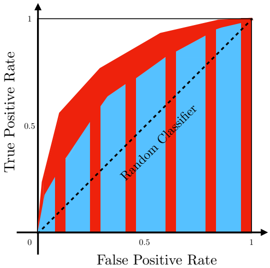
Learning from Imbalanced Data
Impact of Imbalance

Example of linear SVM and k-NN with 50% and 20% of positives.



Use appropriate measures





$$F_\beta = \frac{(1+\beta^2)(P - FN)}{(1+\beta^2)P - FN + FP}$$

$$\mathbb{P}[f(x_+) > f(x_-)]$$

G-mean

Precision

MCC

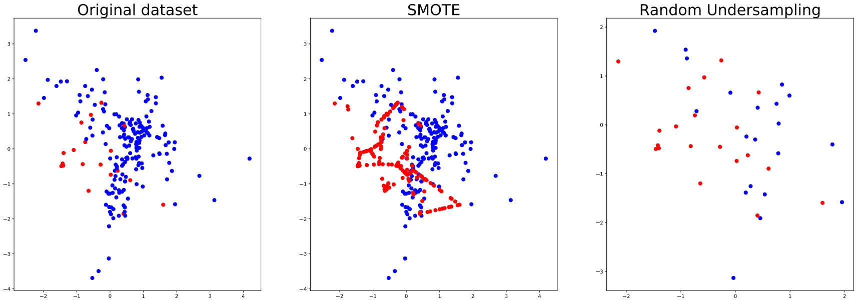
Mean Average Precision

False Positive Rate

Recall

Average Precision

Use sampling strategies

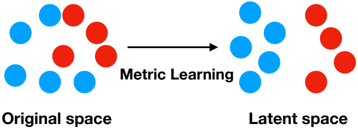


- Oversampling: Random - SMOTE - BorderSMOTE, ...
- Undersampling: Random - Tomek Link - ENN, ...

Distance and representation

$$d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')},$$

where $\mathbf{M}$ is **PSD**.



Cost-sensitive learning

$$C_{TP} = 0 \qquad C_{FN} = c$$

$$C_{FP} = 1 - c \qquad C_{TN} = 0$$

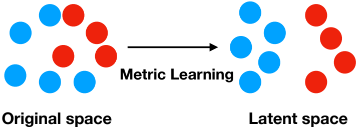
$\rightarrow$ $c \simeq 1$ to encourage low miss-classification on positives.

$$\ell(y, h(\mathbf{x})) = c \cdot y \cdot (1 - h(\mathbf{x})) + (1 - c) \cdot (1 - y) \cdot h(\mathbf{x})$$

Distance and representation

$$d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')},$$

where $\mathbf{M}$ is **PSD**.



→ *Learning Maximum Excluding Ellipsoids from Imbalanced Data with Theoretical Guarantees*, PRL, 2018.

→ *An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data*, ICTAI, 2019.

Cost-sensitive learning

$$C_{TP} = 0 \qquad C_{FN} = c$$

$$C_{FP} = 1 - c \qquad C_{TN} = 0$$

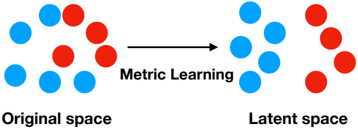
→ $c \simeq 1$ to encourage low miss-classification on positives.

$$\ell(y, h(\mathbf{x})) = c \cdot y \cdot (1 - h(\mathbf{x})) + (1 - c) \cdot (1 - y) \cdot h(\mathbf{x})$$

Distance and representation

$$d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}') = \sqrt{(\mathbf{x} - \mathbf{x}')^T \mathbf{M} (\mathbf{x} - \mathbf{x}')},$$

where $\mathbf{M}$ is **PSD**.



→ *Learning Maximum Excluding Ellipsoids from Imbalanced Data with Theoretical Guarantees*, PRL, 2018.

→ *An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data*, ICTAI, 2019.

Cost-sensitive learning

$$C_{TP} = 0 \qquad C_{FN} = c$$

$$C_{FP} = 1 - c \qquad C_{TN} = 0$$

→ $c \simeq 1$ to encourage low miss-classification on positives.

$$\ell(y, h(\mathbf{x})) = c \cdot y \cdot (1 - h(\mathbf{x})) + (1 - c) \cdot (1 - y) \cdot h(\mathbf{x})$$

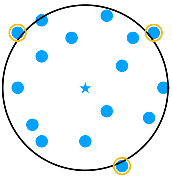
→ *From Cost-Sensitive Classification to Tight F-Measure Bounds*, AISTATS, 2019.

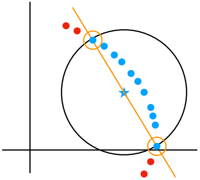
→ *Tree-based Cost Sensitive Methods for Fraud Detection in Imbalanced Data* , IDA, 2018.

Outline

Frauds are close to each other, they form small groups in the feature space

Given a set of m unlabelled points, find the center $\mathbf{c}$ and the **smallest** radius R of the ball that includes the data (Tax and Duin, 2004).

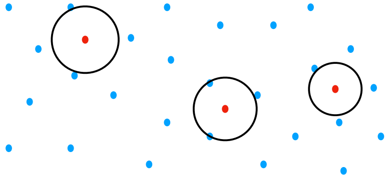




$$\min_{\mathbf{c},R,\xi} \quad R^2 + \frac{\mu}{m}\sum_{i=1}^m \xi_i, \quad \text{s.t.} \quad \|\mathbf{x}_i - \mathbf{c}\|_2^2 \le R^2 + \xi_i,\ \forall i, \quad \xi_i \ge 0\ \forall i.$$

$$\min_{\mathbf{c},\rho,\xi} \quad \frac{1}{2}\|\mathbf{c}\|_2^2 + \frac{1}{\nu m}\sum_{i=1}^m \xi_i - \rho - \frac{1}{2}\|\mathbf{x}_i\|_2^2, \quad \text{s.t.} \quad \mathbf{c}^T\mathbf{x}_i \ge \rho + \frac{1}{2}\|\mathbf{x}_i\ \forall i, \quad \xi_i \ge 0\ \forall i.$$

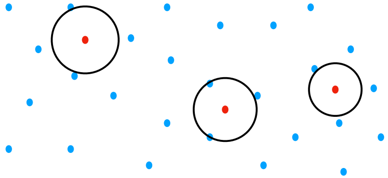
Being in the ball $\iff$ being above the hyperplane



- Use the idea of MIB to create MEB
- One model per positive instance
- Require few positive neighbors

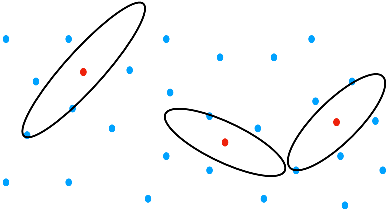
ME^2: Learning Risky Areas

From MIB to ME^2



- Use the idea of MIB to create MEB
- One model per positive instance
- Require few positive neighbors

↓ Learning a Metric



- From balls to ellipsoids
- Increase decision boundary

$\rightarrow$ Maximum Excluding Ellipsoids

ME^2: Learning Risky Areas

Optimization problem

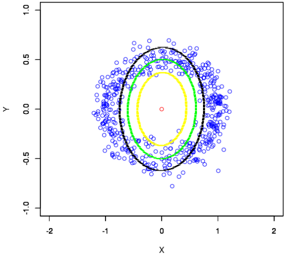
$$\min_{R,\mathbf{M},\boldsymbol{\xi}} \quad \frac{1}{m}\sum_{i=1}^{m}\xi_i + \mu(B-R)^2 + \lambda\|\mathbf{M}-\mathbf{I}\|_{\mathcal{F}^2},$$

$$s.t. \quad \|\mathbf{x}_i - \mathbf{c}\|_{\mathbf{M}}^2 \geq R - \xi_i, \quad \forall i = 1,\dots,m,$$

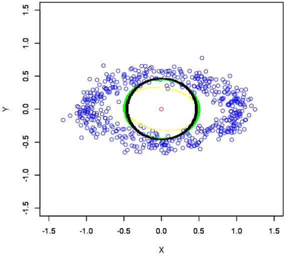
$$\xi_i \geq 0, \quad \forall i = 1,\dots,m$$

$$0 \leq R \leq B,$$

error terms (in terms of distances) regularization term



Influence of μ



Influence of λ

ME^2: Learning Risky Areas

Dual formulation

→ express the Lagrangian $\mathcal{L}$ including the constraints
→ expression of primal variables w.r.t. dual ones:

1. derivative of $\mathcal{L}$ w.r.t. primal variables
2. set derivatives to 0

ME^2: Learning Risky Areas

Dual formulation

$\rightarrow$ express the Lagrangian $\mathcal{L}$ including the constraints
$\rightarrow$ expression of primal variables w.r.t. dual ones:

1. derivative of $\mathcal{L}$ w.r.t. primal variables
2. set derivatives to 0

One of these derivatives gives:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{M}} = 0 \implies \mathbf{M} = \mathbf{I} + \frac{1}{2\lambda} \sum_{i=1}^{m} \alpha_k (\mathbf{x}_k - \mathbf{c})(\mathbf{x}_k - \mathbf{c})^T.$$

$\rightarrow$ M **is Positive Semi Definite for free**

Using stability framework (Bousquet and Elisseeff, 2002)

$$\mathcal{R}(\mathbf{M}, R) \leq \mathcal{R}_S(\mathbf{M}, R) + \mathcal{O}\left(\frac{1}{\min(\mu, \lambda)}\sqrt{\frac{\ln(1/\delta)}{2m}}\right),$$

where $\mathcal{R}_S(\mathbf{M}, R) = \dfrac{1}{m}\sum_{i=1}^{m}[R - \|\xi - \mathbf{c}\|_{\mathbf{M}}^2]_+$.

Using stability framework (Bousquet and Elisseeff, 2002)

$$\mathcal{R}(\mathbf{M}, R) \leq \mathcal{R}_S(\mathbf{M}, R) + \mathcal{O}\left(\frac{1}{\min(\mu, \lambda)}\sqrt{\frac{\ln(1/\delta)}{2m}}\right),$$

where $\mathcal{R}_S(\mathbf{M}, R) = \dfrac{1}{m}\sum_{i=1}^{m}[R - \|\xi - \mathbf{c}\|^2_{\mathbf{M}}]_+$.

the true risk on the underlying and unknown distribution

Using stability framework (Bousquet and Elisseeff, 2002)

$$\mathcal{R}(\mathbf{M}, R) \leq \mathcal{R}_S(\mathbf{M}, R) + \mathcal{O}\left(\frac{1}{\min(\mu, \lambda)}\sqrt{\frac{\ln(1/\delta)}{2m}}\right),$$

where $\mathcal{R}_S(\mathbf{M}, R) = \dfrac{1}{m}\sum_{i=1}^{m}[R - \|\xi - \mathbf{c}\|_{\mathbf{M}}^2]_+$.

the true risk on the underlying and unknown distribution

the empirical risk over the sample S

Using stability framework (Bousquet and Elisseeff, 2002)

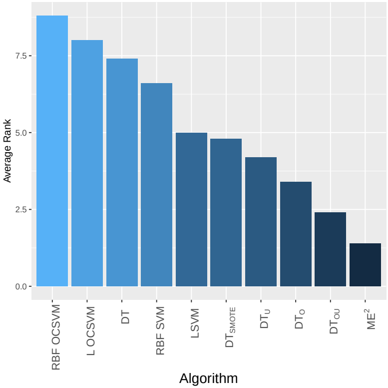
$$\mathcal{R}(\mathbf{M}, R) \leq \mathcal{R}_S(\mathbf{M}, R) + \mathcal{O}\left(\frac{1}{\min(\mu, \lambda)}\sqrt{\frac{\ln(1/\delta)}{2m}}\right),$$

where $\mathcal{R}_S(\mathbf{M}, R) = \dfrac{1}{m}\sum_{i=1}^{m}[R - \|\xi - \mathbf{c}\|^2_{\mathbf{M}}]_+$.

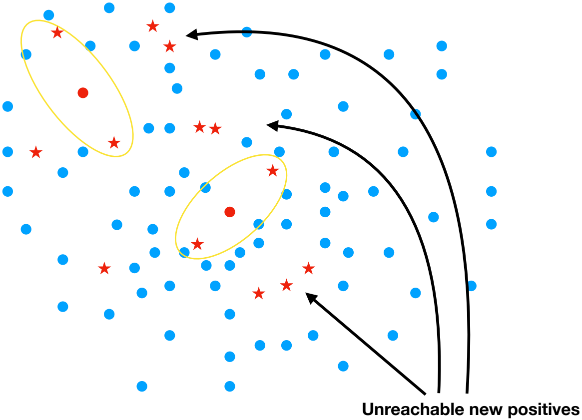
- the true risk on the underlying and unknown distribution
- the empirical risk over the sample S
- generalization gap of the learned model: depends on the complexity of the model

Comparison with standards algorithms on imbalanced datasets

Dataset	Nb. of ex.	% Pos.
Wine	1 599	3.3
Abalone17	2 338	2.5
Yeast6	1 484	2.4
Abalone20	1 916	1.4
Blitz	15 000	1.0



Lower Rank: able to reach better performance



Find a way to increase the influence zone of positives

Outline

Observations

Imbalanced setting $\rightarrow$ low density of positives

low density of positives $\rightarrow$ small influence area

Observations

Imbalanced setting $\rightarrow$ low density of positives

low density of positives $\rightarrow$ small influence area

Idea

Bring points closer to positives by modifying their distances

Observations

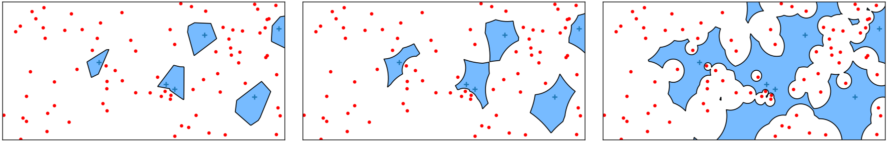
Imbalanced setting $\rightarrow$ low density of positives

low density of positives $\rightarrow$ small influence area

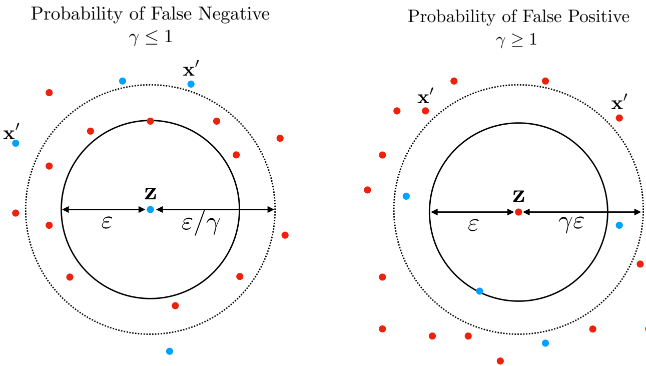
Idea

Bring points closer to positives by modifying their distances

$$d_\gamma(\mathbf{x}, \mathbf{x}_i) = \begin{cases} d(\mathbf{x}, \mathbf{x}_i) & \text{if } y_i = -1, \\ \gamma \cdot d(\mathbf{x}, \mathbf{x}_i) & \text{if } y_i = +1. \end{cases}$$



Importance of the parameter γ



$$FN_\gamma(\mathbf{z}) = \left(1 - \mathbb{P}\left(\mathbf{x}' \in \mathcal{S}_{\frac{\varepsilon}{\gamma}}\right)\right)^{m_+} \qquad FP_\gamma(\mathbf{z}) = \left(1 - \mathbb{P}\left(\mathbf{x}' \in \mathcal{S}_{\varepsilon\gamma}\right)\right)^{m_-}$$

Choose $\gamma < 1$ in Imbalanced settings

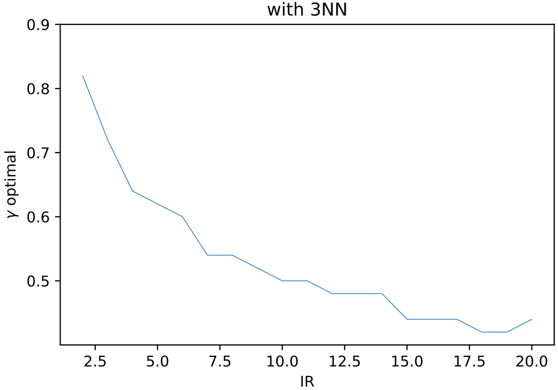
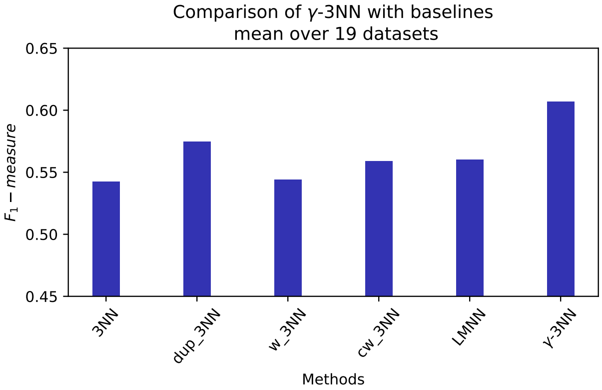


Illustration of the optimal γ with respect to the IR on Balance dataset

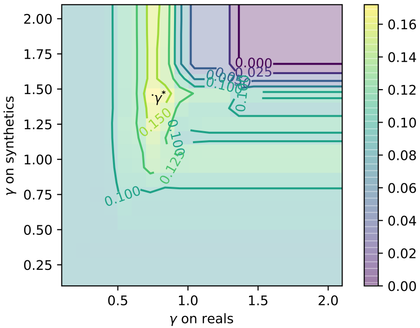
γ-k-NN: a revisit of the k-NN

Experiemental results 1/2

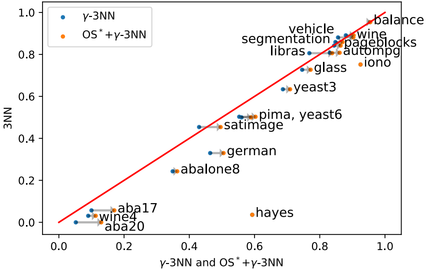


$\rightarrow$ perform better and even better than a Metric Learning approach.

Behaviour of γ-k-NN combined with an over-sampler.



reals: $\gamma < 1$
synthetics: $\gamma > 1$



Coupling with sampling strategies improves the algorithm

Outline

Objective: find a way to optimize the F-measure F_β

$$F_\beta = \frac{(1+\beta^2)(P-FN)}{(1+\beta^2)P - FN + FP} = \frac{(1+\beta^2)(P-e_1)}{(1+\beta^2)P - e_1 + e_2}.$$

Two important quantities: $e_1 = FN$ et $e_2 = FP$ linked to the empirical risk $\mathcal{R}$.

Objective: find a way to optimize the F-measure F_β

$$F_\beta = \frac{(1+\beta^2)(P-FN)}{(1+\beta^2)P - FN + FP} = \frac{(1+\beta^2)(P-e_1)}{(1+\beta^2)P - e_1 + e_2}.$$

Two important quantities: $e_1 = FN$ et $e_2 = FP$ linked to the empirical risk $\mathcal{R}$.

How to make the link between F_β and $\mathcal{R}$?

Objective: find a way to optimize the F-measure F_β

$$F_\beta = \frac{(1+\beta^2)(P-FN)}{(1+\beta^2)P-FN+FP} = \frac{(1+\beta^2)(P-e_1)}{(1+\beta^2)P-e_1+e_2}.$$

Two important quantities: $e_1 = FN$ et $e_2 = FP$ linked to the empirical risk $\mathcal{R}$.

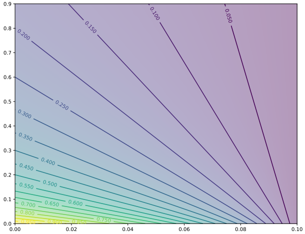
How to make the link between F_β and $\mathcal{R}$?

$\rightarrow$ **Pseudo linearity of the F-measure !**

- Based on previous work published in 2014 at NIPS (Parambath et al., 2014)
- Use the pseudo-linearity of the F-measure
- Derive bounds on optimality of F_β
- Algorithmitic : grid approach

- Based on previous work published in 2014 at NIPS (Parambath et al., 2014)
- Use the pseudo-linearity of the F-measure
- Derive bounds on optimality of F_β
- Algorithmitic : grid approach

$\rightarrow$ **Extend the existing work from both theoretical and practical aspect**



- F_β level sets are hyperplanes in the (e_1, e_2)-space:

$$\forall t \in [0,1],\ F_\beta(\mathbf{e}) = t \iff \exists\, \mathbf{a}, b \text{ t.q. } \langle \mathbf{a}(t), \mathbf{e} \rangle + b(t) = 0.$$

- $\mathbf{a}$: weights assigned to the errors
- $\langle \mathbf{a}(t), \mathbf{e} \rangle$: weighted version of $\mathcal{R}$.

$\rightarrow$ Good choice of $t \iff$ Optimizing F_β.

- Write the difference of F-measures between $\mathbf{e}$ and $\mathbf{e}'$

$$F(\mathbf{e}') - F(\mathbf{e}) = \Phi_{\mathbf{e}} \cdot \langle \mathbf{a}(F(\mathbf{e}'), \mathbf{e} - \mathbf{e}' \rangle, \ \Phi_{\mathbf{e}} = \frac{1}{(1+\beta^2)P - e_1 + e_2}.$$

- Bound this difference using:
  1. linearity of the inner product
  2. sub-optimality ε_1 of the learned hypothesis

$$\boxed{F(\mathbf{e}') - F(\mathbf{e}) \leq \Phi_{\mathbf{e}}\varepsilon_1 + \Phi_{\mathbf{e}} \cdot (e_2 - e_1 - (e'_2 - e'_1))\,(t' - t).}$$

Problem: $\mathbf{e}(t') = \mathbf{e}' = (e'_1, e'_2)$ is unknown

$\rightarrow$ Bound the difference $e'_2 - e'_1$

- When $t' < t$:

$$M_{max} = \max_{\substack{\mathbf{e}'' \in \mathcal{E}(\mathcal{H}) \\ s.t.\ F(\mathbf{e}'') > F(\mathbf{e})}} (e''_2 - e''_1).$$

$$F(\mathbf{e}') \leq F(\mathbf{e}) + \Phi_{\boldsymbol{e}} \varepsilon_1 + \Phi_{\boldsymbol{e}} \cdot (e_2 - e_1 - M_{\max})(t' - t),$$

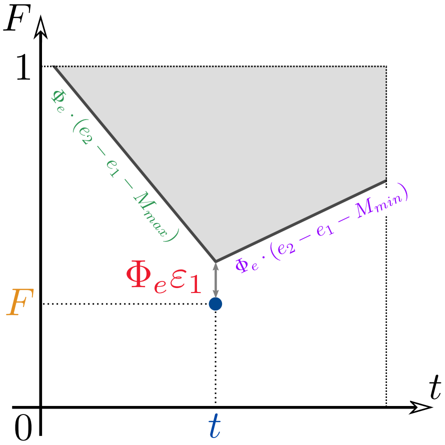
- When $t' > t$:

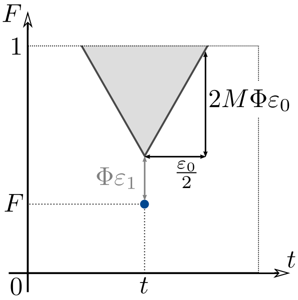
$$M_{\min} = \min_{\substack{\mathbf{e}'' \in \mathcal{E}(\mathcal{H}) \\ s.t.\ F(\mathbf{e}'') > F(\mathbf{e})}} (e''_2 - e''_1).$$

$$F(\mathbf{e}') \leq F(\mathbf{e}) + \Phi_{\boldsymbol{e}} \varepsilon_1 + \Phi_{\boldsymbol{e}} \cdot (e_2 - e_1 - M_{min})(t' - t),$$

CONE: an Algorithm for F-measure Optimization

An asymmetric cone





Interpretation existing bound

of Parambath et al. (2014)

$$F(\mathbf{e}') \leq F(\mathbf{e}) + \Phi \cdot (2\varepsilon_0 M + \varepsilon_1)$$

$$F(\mathbf{e}') \leq F(\mathbf{e}) + \Phi\varepsilon_1 + 4M\Phi|t' - t|.$$

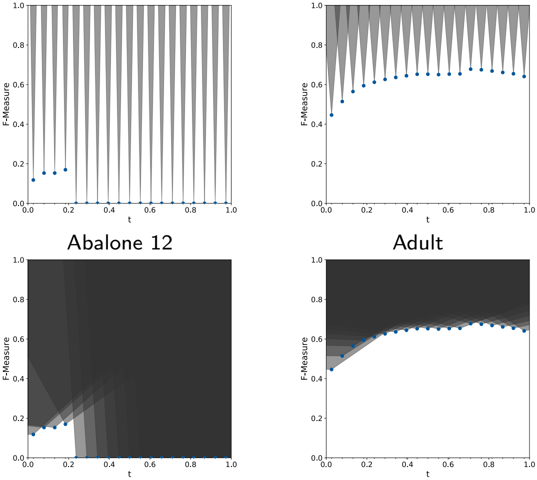
où

- ε_0: distance to optimal weights
- $M = \max_{\mathbf{e}''} \|\mathbf{e}''\|_2$
- $\Phi = (\beta^2 P)^{-1}$ independent from $\mathbf{e}$

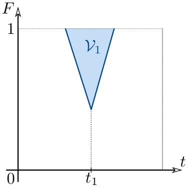
CONE: an Algorithm for F-measure Optimization

Bounds comparison: Parambath et al. (2014) vs Our

Similar bounds ...



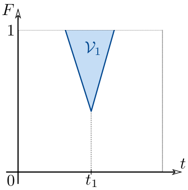
... with highly different slopes.



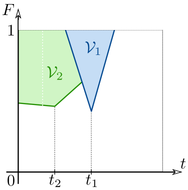
Step 1

Step 1: Take the middle of the t-space of reasearch: $t_1 = 0.5$

$\rightarrow$ Highest values of F in $[0, t_1]$



Step 1



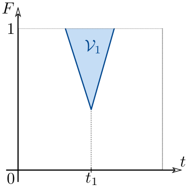
Step 2

Step 1: Take the middle of the t-space of reasearch: $t_1 = 0.5$

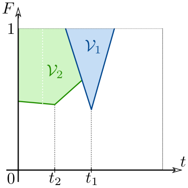
$\rightarrow$ Highest values of F in $[0, t_1]$

Step 2: Choose t_2 in the middle of $[0, t_1]$

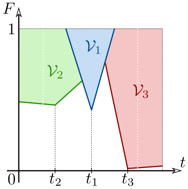
$\rightarrow$ Highest values of F in $[t_1, 1]$



Step 1



Step 2



Step 3

Step 1: Take the middle of the t-space of reasearch: $t_1 = 0.5$

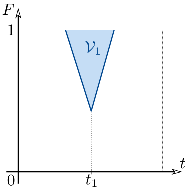
$\rightarrow$ Highest values of F in $[0, t_1]$

Step 2: Choose t_2 in the middle of $[0, t_1]$

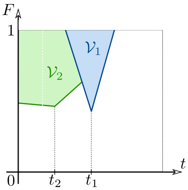
$\rightarrow$ Highest values of F in $[t_1, 1]$

Step 3: Choose t_3 in the middle of $[t_1, 1]$

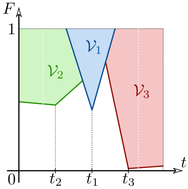
$\rightarrow$ Highest values of F in $[t_1, t_3]$



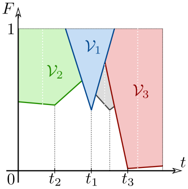
Step 1



Step 2



Step 3



Step 4

Step 1: Take the middle of the t-space of reasearch: $t_1 = 0.5$

$\rightarrow$ Highest values of F in $[0, t_1]$

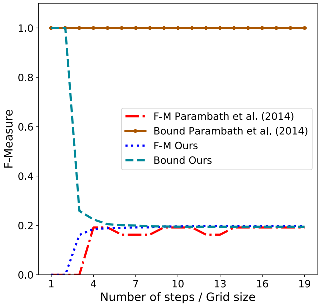
Step 2: Choose t_2 in the middle of $[0, t_1]$

$\rightarrow$ Highest values of F in $[t_1, 1]$

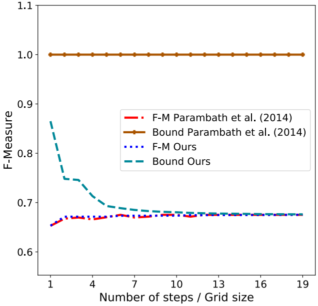
Step 3: Choose t_3 in the middle of $[t_1, 1]$

$\rightarrow$ Highest values of F in $[t_1, t_3]$

Step 4: Choose t_4 in the middle of $[t_1, t_3]$



Abalone 12

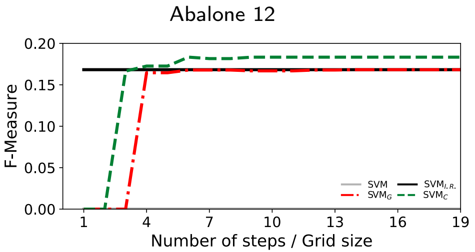


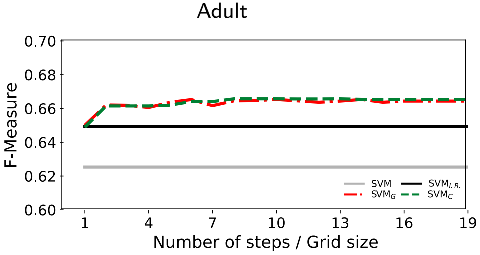
Adult

→ A more informative bound

→ A faster convergence

→ Improves the performances





SVM: a linear SVM

SVM$_{I.R}$: a linear SVM with weighted errors

SVM$_G$: grid approach (Parambath et al., 2014)

SVM$_C$: our approach

Outline

Currently

- Model based on classification error (Decision Tree (Breiman et al., 1984) and Giny criterion)
- Limits the number of alarms
- Focuses on the number of false alarms, i.e. high precision

→ **Does not take main criterion into account: benefits**

Currently

- Model based on classification error (Decision Tree (Breiman et al., 1984) and Giny criterion)
- Limits the number of alarms
- Focuses on the number of false alarms, i.e. high precision

→ **Does not take main criterion into account: benefits**

Idea

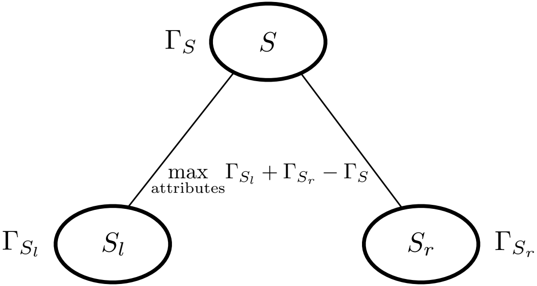
- Define a new loss which optimizes retailers benefits
- Use the amount in the loss function

Compute retailers benefits using a cost matrix (Elkan, 2001)

	Predicted Positive	Predicted Negative
Actual Positive	C_{TP}	C_{FN}
Actual Negative	C_{FP}	C_{TN}

$$C_{TP} = 0 \qquad C_{FN} = (r - c(m)) \cdot m$$
$$C_{FP} = \rho \cdot r \cdot m - \xi \qquad C_{TN} = r \cdot m$$

$$\ell(y, \hat{y}) = \sum_{i=1}^{m} \left[y_i(\hat{y}_i c_{TP_i} + (1 - \hat{y}_i) c_{FN_i}) + (1 - y_i)(\hat{y}_i c_{FP_i} + (1 - \hat{y}_i) c_{TN_i}) \right].$$



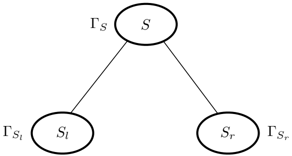
Decision tree

Impurity: $\Gamma = 1 - \sum_{i \in \mathcal{Y}} p_i^2$

Split: $\max\limits_{attributes} \sum_{v \in \text{Children}} \Gamma_S - \alpha_v \Gamma_{S_v}$

Weighted version

$$\Gamma_S = \frac{1}{m} \sum_{i \in S_-} \left[\frac{m_+}{m} c_{FP_i} + \frac{m_-}{m} c_{TN_i} \right] + \frac{1}{m} \sum_{i \in S_+} \left[\frac{m_+}{m} c_{TP_i} + \frac{m_-}{m} c_{FN_i} \right]$$

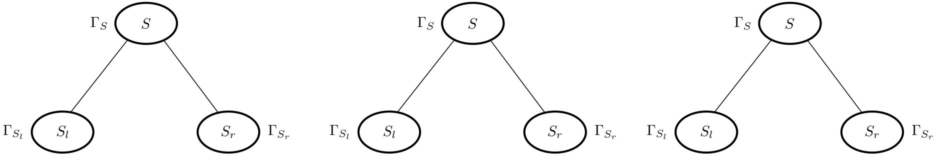


Label

Choose the label that maximizes to profits

Improving Retailers Benefits

Decision tree and splitting criterion 2/2



Label

Choose the label that maximizes to profits

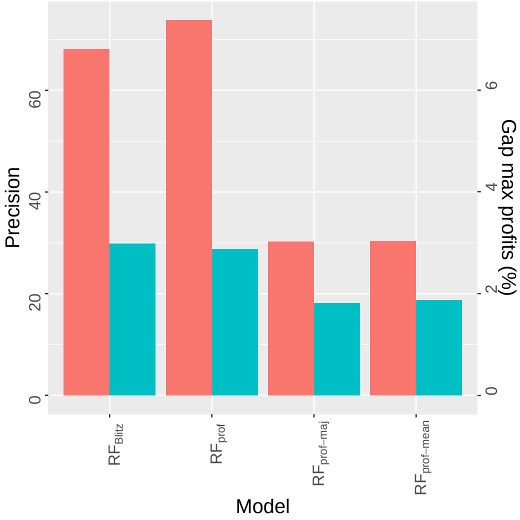
Random Forest

Build several decision trees using the splitting criterion
Combination using different rules:

- simple majority vote
- weighted majority vote using the induced benefits

Improving Retailers Benefits

Experiments



4 months of transactions:

- Improves the profits
- Reduces the precision

A gap of 1% represents around 43 000 euros.

Boosting: Combine models such that f_t compensates for F_{t-1} weaknesses.

$$F_T = f_0 + \sum_{t=1}^{T} \alpha_t f_t$$

Gradient Boosting: Same idea, but work in the prediction space rather than parameter space.

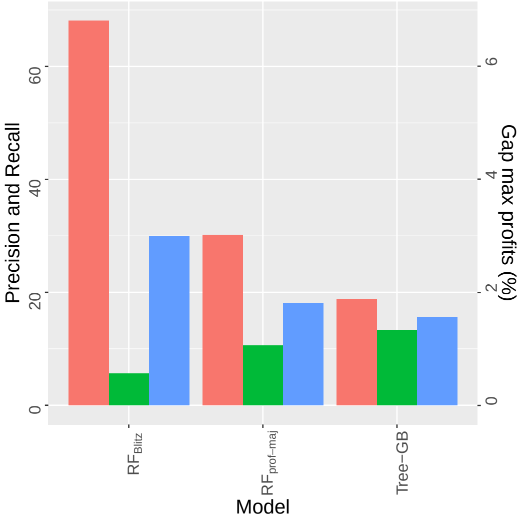
$$g_t = -\left[\frac{\partial \ell(y, F_{t-1}(\mathbf{x}_i))}{\partial F_{t-1}(\mathbf{x}_i)}\right], \quad (f_t, \alpha_t) = \underset{\alpha,\, f}{\text{argmin}} \sum_{i=1}^{m} (r_i - \alpha f(x_i))^2.$$

$\rightarrow$ Give a surrogate of predefined ℓ using the exponential

$$\ell(\mathbf{x}_i, y_i) = y_i(1 - c_i)\exp(-F(\mathbf{x}_i)) + c_i(1 - y_i)\exp(F(\mathbf{x}_i)).$$

Improving Retailers Benefits

Experiments



Using Gradient Boosting

- Reduces training process
- Improves profits
- Higher recall
- Lower precision

Save around 60 000 euros compared to the current one

Outline

Two main axes were proposed to deal with the problem of learning from imbalanced data:

Two main axes were proposed to deal with the problem of learning from imbalanced data:

1. Geometric: based on the distance to positives
   - Risky areas + local learning
   - Modification of the k-NN, modifying distance to positives

Two main axes were proposed to deal with the problem of learning from imbalanced data:

1. Geometric: based on the distance to positives
    - Risky areas + local learning
    - Modification of the k-NN, modifying distance to positives
2. Cost-Sensitive Learning: Weighting the errors
    - Bounds + iterative algorithm: optimizing F-measure
    - Loss + algorithm: improving retailers benefits

	Advantages	**Disadvantages**
$\boldsymbol{ME^2}$	Easy to learn $\mathbf{M}$ Theoretical guarantees on FP	Over-fitting Detect new positives
γ-**k-NN**	Easy to implement Simplicity	Distance Computation Too simple
$\mathbf{CONE}$	Bounds on F_β Derivation of an algorithm Require only few iterations	Algorithm convergence Guarantee at test time
$\boldsymbol{GB}_{\text{Tree}}$	Fast to learn Flexibility	Low Precision

Based on the work on LMNN Weinberger and Saul (2009)

$\rightarrow$ Propose a version of γ-k-NN based on learning new representations.

Ideas :

- Keep compromise FN vs FP.
- Hyper-parameters: optimized to maximize the F-measure

Deriving theoretical guarantees:

- On the learned metric (Bellet et al., 2015)
- On the classification performances

$\rightarrow$ Ongoing work : submission at AISTATS 2020

Lemma: The difference $(e_1 - e_2)$ is a decreasing function of t when $\mathbf{e}(t)$ is obtained from an optimal classifier h learned with the weights $\mathbf{a}(t)$.

Lemma: The difference $(e_1 - e_2)$ is a decreasing function of t when $\mathbf{e}(t)$ is obtained from an optimal classifier h learned with the weights $\mathbf{a}(t)$.

Example when $t' > t$:

$$F(\mathbf{e}') - F(\mathbf{e}) = \Phi_{\mathbf{e}}\left(\underbrace{\langle \mathbf{a}(t), \mathbf{e}\rangle} + (t'-t)(e_2 - e_1) - \underbrace{\langle \mathbf{a}(t'), \mathbf{e}'\rangle}\right),$$

$$= \Phi_{\mathbf{e}}\left(\underbrace{t(e_2 - e_1)} + \underbrace{(1+\beta^2)e_1 - (1+\beta^2)e_1'} - t'(e_2' - e_1') + (t'-t)(e_2 - e_1)\right),$$

$\downarrow$ Use of the Lemma

$$\geq \Phi_{\mathbf{e}}\left(\underbrace{t(e_2' - e_1') - t'(e_2' - e_1')} + (1+\beta^2)(e_1 - e_1') + (t'-t)(e_2 - e_1)\right),$$

$\downarrow$ ...

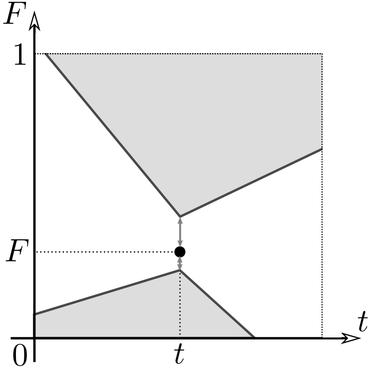
$$F(\mathbf{e}') - F(\mathbf{e}) \geq \Phi_{\mathbf{e}}\left((1+\beta^2)(e_1 - e_1') + (t'-t)e_2 - e_1 - (e_2' - e_1')\right)).$$

$$F(\mathbf{e}') - F(\mathbf{e}) \geq \quad \Phi_{\mathbf{e}}\left((1+\beta^2)(e_1 - e_1') + (t'-t)e_2 - e_1 - (e_2' - e_1')\right)$$

- $e_2' - e_1'$: as seen previously.
- $e_1 - e_1'$: find a tight lower-bound.

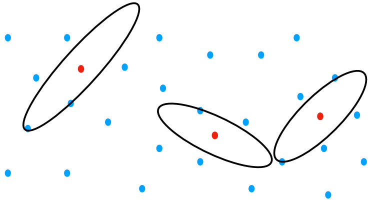
$$F(\mathbf{e}') - F(\mathbf{e}) \geq \quad \Phi_{\mathbf{e}}\left((1+\beta^2)(e_1 - e'_1) + (t'-t)e_2 - e_1 - (e'_2 - e'_1)\right)$$

- $e'_2 - e'_1$: as seen previously.
- $e_1 - e'_1$: find a tight lower-bound.



- Bound the values of F_β
- Get a new algorithm
- Deriving generalization bounds F_β
- Optimality of F_β at test time
- Empirically : generalization bounds based on the **validation** (Kawaguchi et al., 2017).

Problem: ME^2 is prone to over-fitting



$\rightarrow$ Find a way to "smooth" the classification process

$\rightarrow$ Convex Combinations of local models (Zantedeschi et al., 2016)

Thank you for your attention !

International Journal

- G.Metzler, X.Badiche, B.Belkasmi, E.Fromont, A.Habrard and M.Sebban; *Learning Maximum Excluding Ellipsoids from Imbalanced Data with Theoretical Guarantees*, PRL, 2018.

International Conferences

- R.Viola, R.Emonet, A.Habrard, G.Metzler, S. Riou and M.Sebban; *An Adjusted Nearest Neighbor Algorithm Maximizing the F-Measure from Imbalanced Data*, ICTAI, 2019.
- K.Bascol, R.Emonet, E.Fromont, A.Habrard, G.Metzler and M.Sebban; *From Cost-Sensitive Classification to Tight F-Measure Bounds*, AISTATS, 2019.
- G.Metzler, X.Badiche, B.Belkasmi, E.Fromont, A.Habrard and M.Sebban; *Tree-based Cost Sensitive Methods for Fraud Detection in Imbalanced Data* , IDA, 2018.

National Conferences

- R.Viola, R.Emonet, A.Habrard, G.Metzler, S.Riou and M.Sebban; *Une version corrigée de l'algorithme des plus proches voisins pour l'optimisation de la F-mesure dans un contexte déséquilibré*, CAp, 2019.
- K.Bascol, R.Emonet, E.Fromont, A.Habrard, G.Metzler and M.Sebban; *Un algorithme d'optimisation de la F-Mesure par pondération des erreurs de classification*, CAp, 2018.
- G.Metzler, X.Badiche, B.Belkasmi, E.Fromont, A.Habrard and M.Sebban; *Apprentissage de Sphères Maximales d'exclusion avec Garanties Théoriques*, CAp, 2017.

References I

Bellet, A., Habrard, A., and Sebban, M. (2015). Metric learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 9(1):1–151.

Bousquet, O. and Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research*, 2:499–526.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and regression trees*. The Wadsworth statistics/probability series. Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA.

Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, pages 973–978, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Kawaguchi, K., Kaelbling, L. P., and Bengio, Y. (2017). Generalization in deep learning. *arXiv preprint arXiv:1710.05468*.

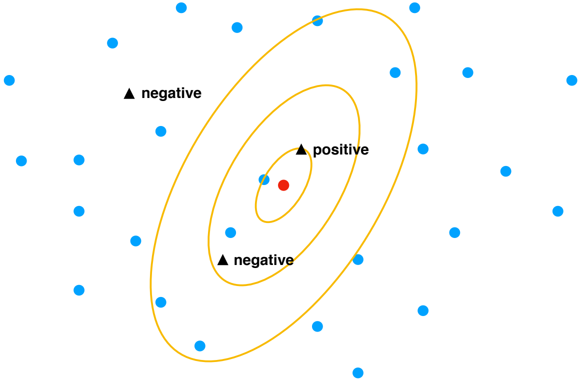
Parambath, S. P., Usunier, N., and Grandvalet, Y. (2014). Optimizing f-measures by cost-sensitive classification. In *Advances in Neural Information Processing Systems (NIPS-14)*, pages 2123–2131.

Tax, D. M. J. and Duin, R. P. W. (2004). Support vector data description. *Machine Learning Journal*, 54(1):45–66.

Weinberger, K. Q. and Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10:207–244.

Zantedeschi, V., Emonet, R., and Sebban, M. (2016). Metric learning as convex combinations of local models with generalization guarantees. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

1. assign each text example to its closest positive
2. apply the following classification rule:



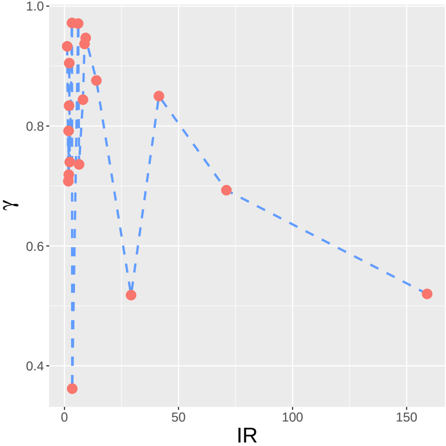
Based on the work on LMNN Weinberger and Saul (2009)

$\rightarrow$ Propose a version of γ-k-NN based on learning new representations.

$$\min_{\mathbf{M}\in\mathbb{S}^+} \quad \frac{1}{m^3}\left(\frac{1-\alpha}{2}\sum_{\substack{\mathbf{x}_i,\mathbf{x}_j\in\mathcal{S}\\ y_i=y_j=1}} d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)^2 + \frac{1-\alpha}{2}\sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)\in\mathcal{R}\\ y_i=1}} \left[1-m'+d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_j)^2 - d(\mathbf{x}_i,\mathbf{x}_k)^2\right]_+ + \alpha\sum_{\substack{(\mathbf{x}_i,\mathbf{x}_j,\mathbf{x}_k)\in\mathcal{R}\\ y_i=-1}} \left[1-m'+d(\mathbf{x}_i,\mathbf{x}_j)^2 - d_{\mathbf{M}}(\mathbf{x}_i,\mathbf{x}_k)^2\right]_+\right) \mu\|\mathbf{M}-\mathbf{I}\|_{\mathcal{F}}^2.$$

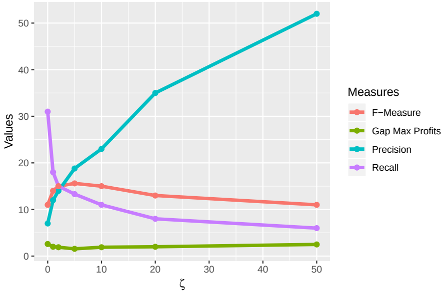
γ-k-NN: a revisit of the k-NN

$\gamma^{\star}$ vs. I.R.



$\rightarrow$ In average γ is a decreasing function of IR

Influence of ξ parameter



Increasing ξ value:

- Improves the Precision
- Reduces the Recall
- Reduces the retailers benefits

Comparison on Blitz dataset

Comparison Contributions